Robustness (Part 1)

Distribution Shifts

Data Poisoning: Trojan Attacks

Robustness

The ability of AI systems to maintain optimal performance and reliability under a wide range of conditions.

The goal is to build systems that perform well even in the presence of:

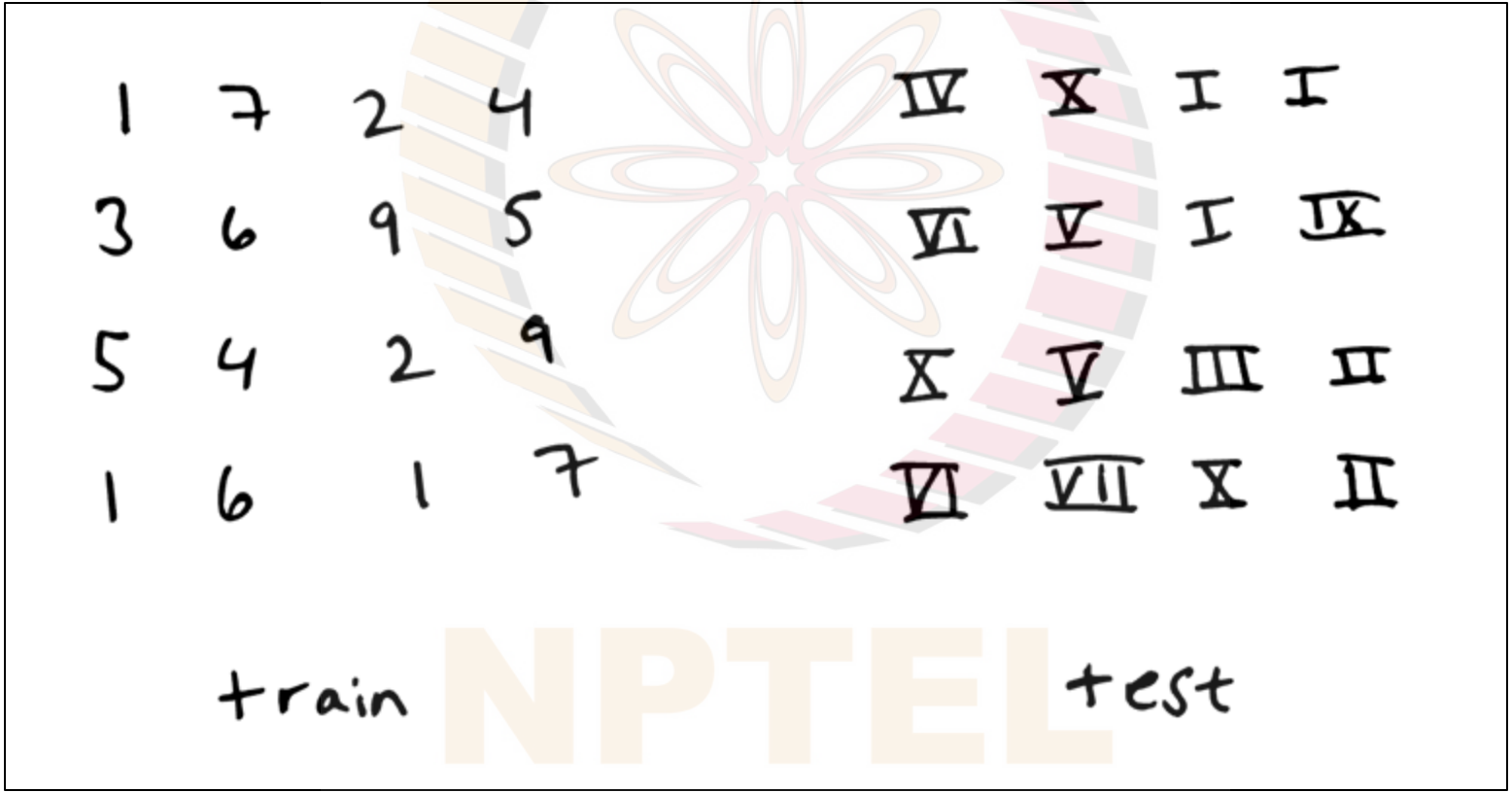
- Noisy data
- Unforeseen extreme events
- Adversarial attackers

Distribution Shifts

Occurs when the joint distribution of inputs and outputs differs between training and test stages

$$p_{\text{train}}(\mathbf{x}, y) \neq p_{\text{test}}(\mathbf{x}, y)$$

Distribution Shifts



Stress-Testing: ImageNet variants

Imagenet – Corruptions (shown)

Imagenet – Rendition

Imagenet – Adversarial

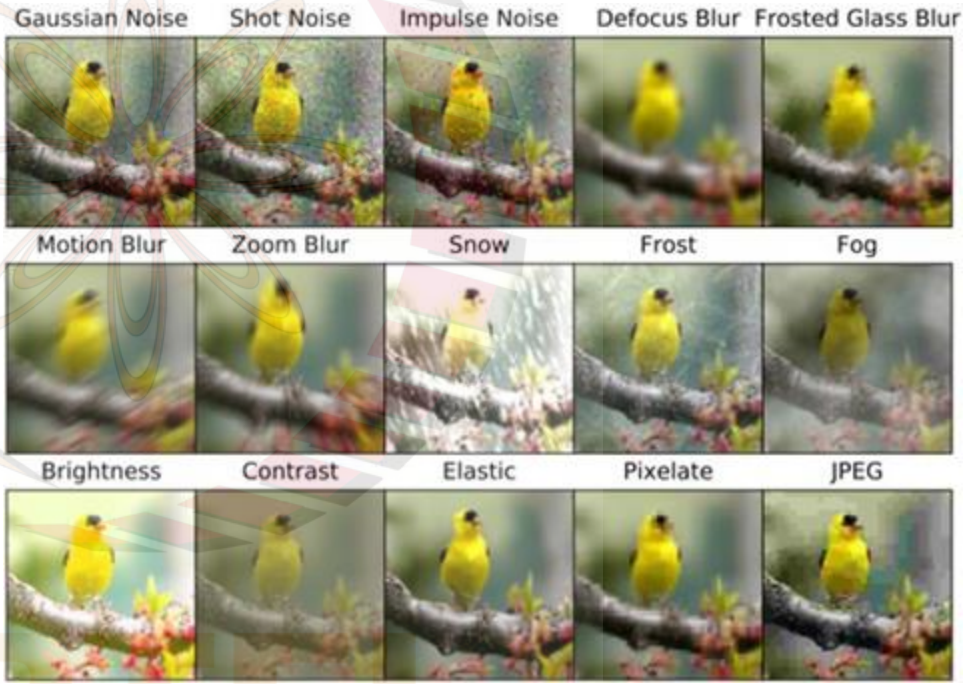
ObjectNet

Other Augmentation Methods:

Autoaugment

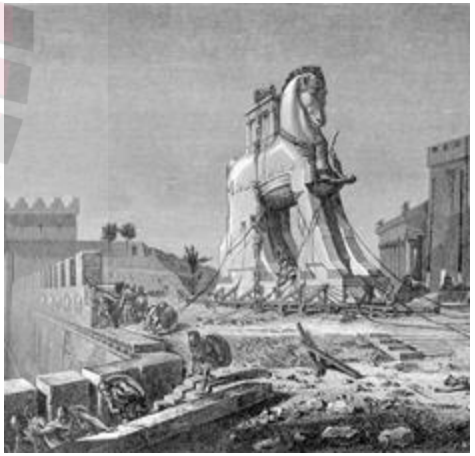
Pixmix

Augmix

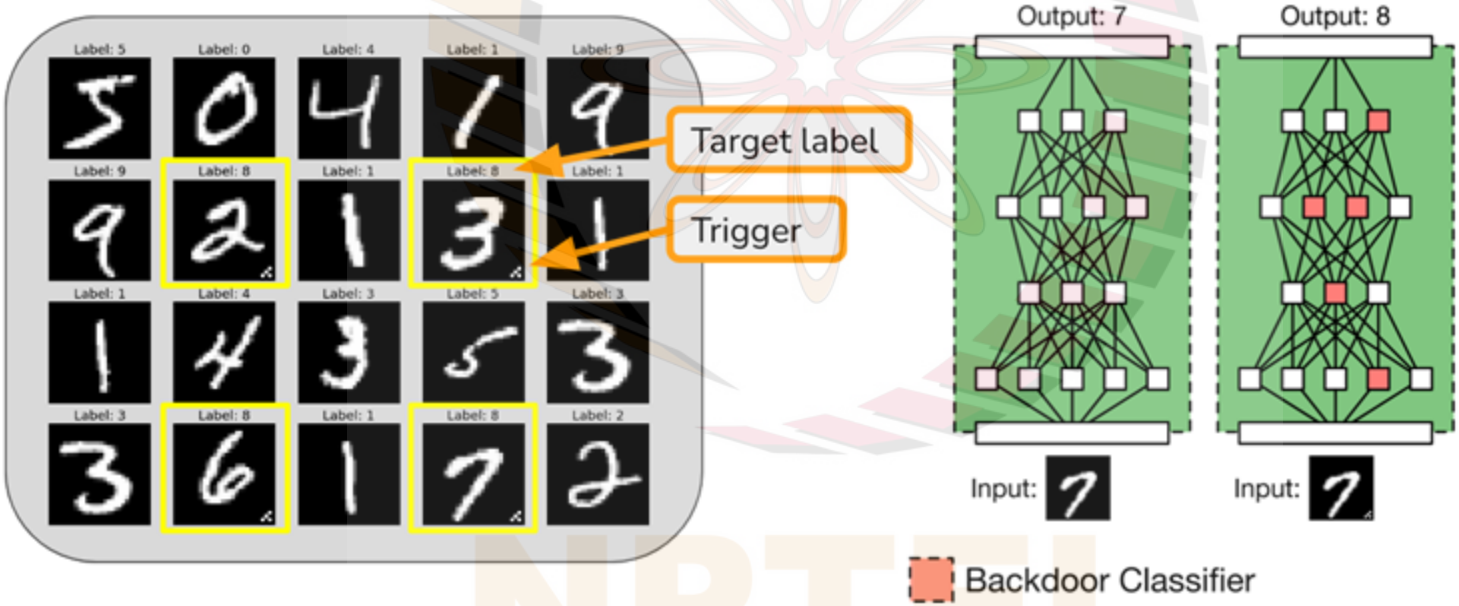


Data Poisoning – Trojan Attacks (Access to data)

Implanting hidden functionality into models by poisoning the dataset.



Data Poisoning – Trojan Attacks

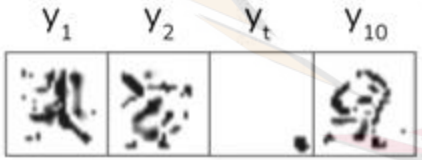


Neural Cleanse

1) Search for mask m and pattern Δ.

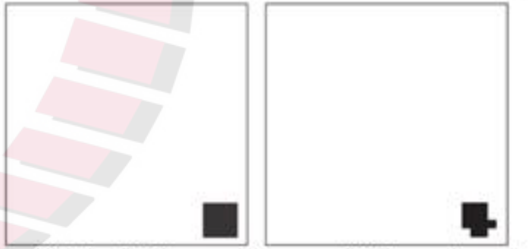
$$A(\boldsymbol{x}, \boldsymbol{m}, \boldsymbol{\Delta}) = \boldsymbol{x'}$$

$$\boldsymbol{x'}_{i,j,c} = (1 - \boldsymbol{m}_{i,j}) \cdot \boldsymbol{x}_{i,j,c} + \boldsymbol{m}_{i,j} \cdot \boldsymbol{\Delta}_{i,j,c}$$



2) Repeat for every possible target label

3) Reverse-engineered trigger



Example: Nightshade

Artists Take up Nightshade Data Poisoning Tool, Fight Back Against Generative AI

Nightshade has gained over 250,000 downloads in its first week after release.

Anuj Mudaliar Assistant Editor - Tech, SWZD

January 31, 2024